2017-05-01

Quantitative Risk Assessment for Residential Mortgages

Qingyun Ren
Worcester Polytechnic Institute

QUANTITATIVE RISK ASSESSMENT FOR RESIDENTIAL MORTGAGES

A Master's Project

Submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirement of the

Degree of Master of Science

In

Financial Mathematics

By

Qingyun Ren

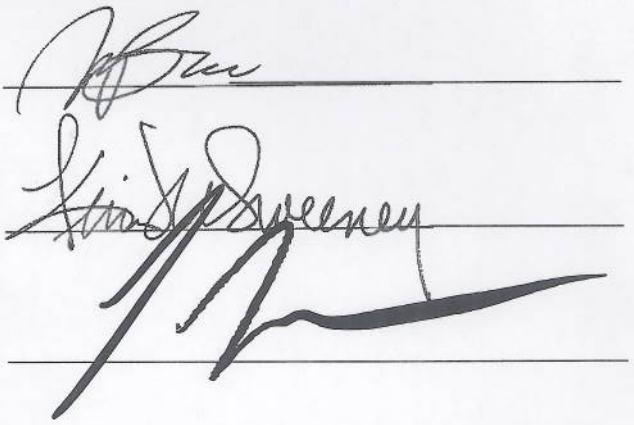
May, 2017

Approved by

Professor Marcel Y. Blais, Project Advisor

Professor Kevin M. Sweeney, Project Advisor

Professor Luca Capogna, Head of Department



QUANTITATIVE RISK ASSESSMENT FOR RESIDENTIAL MORTGAGES

Project Sponsor: EnerScore

Project Advisor: Prof. Marcel Blais

Project Advisor: Prof. Kevin Sweeney

Qingyun Ren, Yi Jiang, Gautami Sant, Sha Fang

ABSTRACT

The crisis of the mortgage market and the mortgage-backed security (MBS) market in 2008 had dramatic negative effects in dragging down all of the economy on a worldwide scale. Many researches have, therefore, attempted to explore the influencing factors on mortgage default risk. This project, in cooperation with the company EnerScore, revolves around discovering a correlation between portfolios of mortgages to underlying energy expenditures. EnerScore's core product provides an internal dataset related to home energy efficiency for American homes and gives their corresponding home energy efficiency rating to every home, which is called an "EnerScore."

This project involves discovering a correlation between default within portfolios of mortgages based on underlying energy expenditures. The goal is to show that energy efficient homes potentially have lower default risks than standard homes because the homes which lack energy efficiency are associated with higher energy costs. This leaves less money to make the mortgage payment, and thereby increases default risk.

The first phase of this project involves finding a foreclosure dataset that will be used to design the quantitative model. Due to limited availability and constraints related to default data, Google search query data is used to develop a broad based and real-time index of mortgage default risk and establish a meaningful scientific correlation.

After analyzing several statistical models to explore this correlation, the regression tree model showed that the EnerScore is a strong predictor for mortgage default risk when using city-level mortgage default risk data and EnerScore data.

Contents

1. EXECUTIVE SUMMARY

Energy represents the third largest cost of housing in the United States, on average. There is increased awareness among homebuyers and mortgage originators about energy efficiency and its importance. This report studies how underlying energy expenditures can result in higher or lower default risk and establishes a correlation between residential mortgage default risk and energy efficiency.

The study highlights the situation for many moderate and middle-income homebuyers and owners. Consider a hypothetical user story of two homeowners from a middle-income segment. Both of these homeowners have similar careers and maintain a similar cost of living standard. One of them enjoys the benefits of a home with energy efficiency, is aware of the efficiency levels, and understands that this efficiency pays for itself over the lifetime of home through lower costs. The other homeowner owns an older home, which lacks energy efficiency upgrades. Upon considering a scenario where oil prices spike and the homeowner's heating bill increases substantially, it becomes more difficult to pay the energy bills. The middle-income homeowner, who does not understand the importance of energy efficiency and is unaware of how the energy efficiency can cut down the potential costs, may default on their mortgage more quickly than the homeowner with lower energy costs.

A quantitative analysis to be discussed later suggests that owners of energy efficient homes potentially have lower default risks than the owners of standard homes, because the homes that lack energy efficiency are associated with higher energy expenditures. This reality is driven by the homeowner's available cash flow and ability to make the mortgage payment.

Reducing this default risk for individual homes may be beneficial to mortgage originators. Further, financial risks are orders of magnitude higher for mortgage-backed securities (MBS) that represent portfolios of loans. Thus, this study examines how a correlation is established between energy expenditures and portfolios of mortgages.

EnerScore is an organization that specializes in property searches tied to the energy efficiency of homes. The EnerScore search tool provides an energy score and estimated energy bill for any home based on home characteristics such as age, size, fuel type, architecture characteristics and sound building science.

The current model of the "EnerScore" has an "A" - "F" rating scale of home energy efficiency available to embed in real estate listings. Its algorithms process tax assessor and building permit data to predict the performance of the building shell and its mechanicals (e.g. power, lighting, heating, etc.). The EnerScore also shows the estimated annual consumption, based on average

U.S. occupancy. It is not affected by habits of the occupant's thermostat setting, lighting use or other equivalent factors.

Home buyers/renters can use the EnerScores to understand the full ownership costs of a home. Home occupants want to get the most out of their home, especially when they sell or rent, and mortgage lenders, like banks, can take into account expected energy bills when they evaluate mortgage risk.

This EnerScore assists in important decision-making about which home to buy or rent by revealing the comparative energy performance – the efficiency and estimated utility costs. This can make a huge difference in long-term affordability. EnerScore can deliver insights into homes' potential for energy performance improvements. If EnerScore's estimated costs are lower than actual expenditures, there are some simple things one can do to lower bills. Lowering thermostat settings, reducing plug loads, changing to high efficiency light bulbs and purchasing Energy Star appliances are some examples.

EnerScore considered that a correlation between energy efficiency and mortgage default risk has been established, and initiated the development of the "EnerScore Portfolio Risk Assessment for Mortgages" (EPRAM). Essentially, this would be a data model that establishes a risk reward ratio akin to a Sharpe ratio [1], but for a portfolio of mortgages (a MBS) based solely on EnerScore's evaluation of energy performance. To financial institutions, the EPRAM would give lenders the ability to manage the risk of portfolios of securitized mortgages. To retail investors, it would provide an indicator of the attractiveness of these instruments.

The Home Energy Rating System (HERS) Index is the industry standard by which a home's energy efficiency is measured. It is also the nationally recognized system for inspecting, testing and calculating a home's energy performance. The current EnerScore model derives a HERS rating estimate based on various sets of public data and records. The outputs of the current EnerScore model are the efficiency levels from A to F, a result of partitioning the 200-point range of positive values into subintervals of size 12.

In order to develop EPRAM, various data sources were under consideration. CoreLogic [2], one of the data sources considered provides information intelligence to identify and manage growth opportunities, improve business performance and manage risk. In addition, CoreLogic provides foreclosure data, containing historical, property-level pre-foreclosure and foreclosure records from initial notice of default to final disposition, judicial and non-judicial foreclosure and foreclosure sales. Due to limited availability of default data, foreclosure data provided by CoreLogic was under consideration. It contains about 46 million records, recorded from the year 2000 to present. The coverage of this data spreads to 85 percent coverage of U.S. foreclosure data. Due to limited availability of access to this data and ongoing evaluation of

issues related to securing data source licensing rights, we utilized an alternative methodology for establishing this correlation.

The paper "Mortgage default risk: New evidence from internet search queries" [3] suggests using Google search query data to develop a broad-based and real-time index of mortgage default risk and prove a scientific correlation. Thus, internet search queries yield valuable new insights into household mortgage default risk after searching queries for terms such as "mortgage assistance" and "foreclosure help" using Google trends and aggregating these Google search queries.

We normalized the amount of queries to comprise MDRI (Mortgage Default Risk Index). By using this approach, we can get the mortgage default risk index for 20 cities in the past one year. This is a city-level EnerScore. Using this approach, we find a correlation between city-level default risk and a city-level EnerScore.

Following this approach, this study finds data in the time range of the last twelve months on Google for search terms like foreclosure, foreclosures, foreclosure assistance, foreclosure help, government assistance mortgage, home mortgage assistance, home mortgage help, housing assistance, mortgage assistance program, mortgage assistance, mortgage foreclosure help, mortgage foreclosure, and mortgage help. These data were analyzed for twenty U.S. cities, including Tampa, Charlotte, Miami, Orlando, Philadelphia, Detroit, Washington, Atlanta, Chicago, New York, Houston, San Antonio, Boston, Denver, Dallas, Austin, San Diego, Phoenix, Seattle, and San Francisco. By using this approach, we were able to create the mortgage default risk index for 20 cities in the past one year. The regression tree model proved that EnerScore is a strong predictor for mortgage default risk by using city-level mortgage default risk data and EnerScore data.

2. ACKNOWLEDGMENTS

We would like to express our sincere gratitude to our sponsor, EnerScore for their active participation, support and resources. We really appreciate the help from Tom Witkin, Brian Butler and Kenn Butler, who provided insightful comments and encouragement that helped us widen our research from various perspectives.

We also want thank our advisors Professor Marcel Blais and Professor Kevin Sweeney for their continuous support, patience, motivation, and immense knowledge. Their guidance helped us significantly, during research and report development.

3. ABOUT ENERSCORE

EnerScore provides an energy score and estimated energy bill for any home, based on home characteristics such as age, size, and sound building science. EnerScore's mission is to promote homes that are more energy efficient and to accelerate adoption of energy efficiency improvements in American homes.

EnerScore aims to help people make informed decisions based on the third largest home expense, energy bills.

3.1 EnerScore Project Objectives

EnerScore's long-term objectives for this project include:

- Deliver a quantitative model that establishes a risk reward ratio for a portfolio of mortgages based solely on energy efficiency.
- Develop an API that integrates this model developed with the EnerScore website.
- Develop recommendations for commercialization of the product

4. DATA MODELING

4.1 Methodology

In statistics regression models and classification models are widely used to solve prediction problems. In this project we adopt linear regression and regression tree methodologies to explore the relationship between mortgage default risk and home energy efficiency.

4.1.1 Multiple Linear Regression

Linear regression is a commonly used modeling technique. Statistically, linear regression is for modeling the relationship between a dependent target variable, Y, and one or more independent variables. Multiple linear regression is to use multiple independent variables, X1, X2, etc. It is different from multivariate linear regression, which returns a set of internal correlated target variables instead of just a single scalar variable for multiple linear regression. Mathematically, the multiple linear regression model is to be expressed as,

$Y_i = \beta_0 1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots$, where Y_i is the target attribute, and β_0, β_1, β_2 … are the coefficients of the independent variables [4].

Given a sample, multiple linear regression usually uses the least-squares estimation method for the estimation of the coefficients. Particularly, in our project we use the ordinary least squares (OLS) technique to minimize the sum of square distance between line in the linear regression model and the sample points. This distance is called the sum of squared residuals.

Multiple linear regression is unbiased and consistent if the residuals error has finite variance and has limited significant correlation with the target attributes. It is considered to be a fundamentally basic algorithm for financial modeling for several reasons.

First, the computational expense for this method is lower than most other algorithms, the residual is easy to calculate, and coefficients are convenient for estimation.

Second, the model is simple. As a linear model, multiple linear regression tends to avoid overfitting. That is also the reason why linear regression is often used as a benchmark in studies. Any fitted model, either a financial model or some other kind, must beat linear regression in accuracy to be justified.

For the above reasons we choose a multiple linear regression model for our analysis.

With these advantages, the downside for the model is also clear. Multiple linear regression may not be the best fit for many datasets, especially complicated ones. On the one hand, multiple linear regression only looks at the linear relationship and mean of the variables. The method does a poor job of accounting for extreme values in the data that fall far from the best-fitting hyperplane. On the other hand, it also depends on the assumption that all predictor variables,

X1, X2, etc., are independent. In reality most of the financial variables are not completely independent.

4.1.2 Regression Tree

Regression tree learning is an implementation of a decision tree. It is a predictive model using observations about an instance, as in branch, to predict the instance's expected value, as in leaf node [5].

The methodology for a regression tree is to split a dataset into smaller and smaller subsets while a top-down tree is built. The step for splitting data is called "branching" and the final split subsets are called "leaf" nodes. For numerical data as we are using in our project, the splitting criteria is called Standard Deviation Reduction (SDR). It is used to calculate the standard deviation of a subset before and after a split. If the standard deviation is reduced, it is suggested to be a good split. SDR records the decrease of all possible splits, and finds the largest of these decreases to make an actual split.

Pang-Ning Tan mentioned a regression tree sample (figure 1) in his book "*Introduction to Data Mining*" [5], which could be a good example. The tree starts from node P, then it divides into smaller subsets Q and R. It finally predicts 0 and 1 under each node.

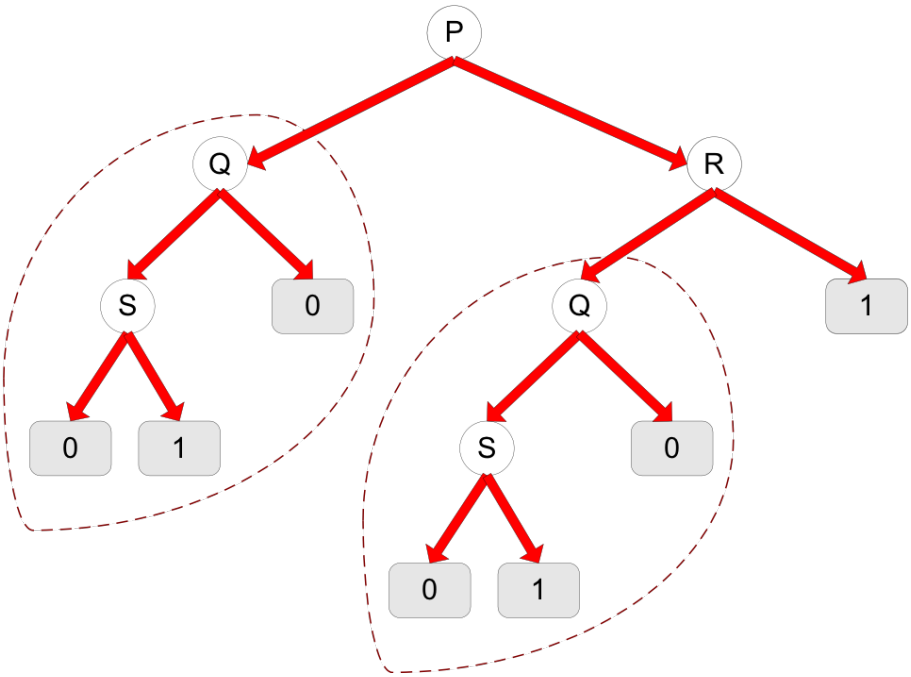


Figure 1: A sample tree in "*Introduction to Data Mining*"

The predictive values are the regression trees' leaf nodes and are calculated by fitting a linear regression model among samples in a particular leaf node, outputting the target value as the output value. If a new dataset (test set) is given, the prediction of each node is calculated using the same linear function.

4.2 Data Description

4.2.1 Home Energy Efficiency

There are 18 city-level data points connecting home energy efficiency with the mortgage default risk index. The home energy efficiency data is provided by EnerScore and contains attributes: *EnerScore, Hotwater, Heating, Baseload, Cooling, Heating_Fuel_Type, DHW_Fuel_Type (Domestic Hot Water), Cooling_Fuel_Type, Baseload_Fuel_Type, Total_Cost.*

We use the mode to represent the city-level situation for each categorical variable (*Heating_Fuel_Type, DHW_Fuel_Type, Cooling_Fuel_Type, Baseload_Fuel_Type*) and the mean for continuous variables (*Hotwater, Heating, Baseload, Cooling, Total_Cost*).

The EnerScore is EnerScore's evaluation toward home energy efficiency of each city based on specific address in the form of alphabetic scores, where A represents the best grade of the most efficient house and F represents the worst. According to the criterion of EnerScore, we map these letters into numbers from 0 to 180. See table 1 for details. In order

A	A-	B+	B	B-	C+	C	C-
0	12	24	36	48	60	72	84
D+	D	D-	E+	E	E-	F+	F
96	108	120	132	144	156	168	180

Table 1: Scoring rating bar of EnerScore

to map one score to each city, we choose either the mean or mode of EnerScore of the entire city, namely *Ener_mean* and *Ener_mode*. By examining these, we aim to figure out which preprocessing method will give us the best prediction. The remaining attributes represent the usage of energy cost. Note that *DHW_Fuel_Type* only contains gas, *Cooling_Fuel_Type only contains electric, as well as Baseload_Fuel_Type*, and they will not make any difference in the results. Hence, we remove these three factors in our experiments. *Heating_Fuel_Type* is a nominal attribute includes electricity, oil & gas. We replace it with numbers (electricity = 0, gas = 1, oil = 2).

4.2.2 Mortgage Default Risk Index (MDRI)

In this paper we use Google search query data to develop a broad-based index of mortgage default risk. Marcelle (2016) showed that the Mortgage Default Risk Index (MDRI) data collected from Google Trends is persuasive in measuring mortgage default risk [2].

We perform Google search queries for the following terms on 20 cites: foreclosure,

foreclosures, foreclosure assistance, foreclosure help, government assistance mortgage, home mortgage assistance, home mortgage help, housing assistance, mortgage assistance program, mortgage assistance, mortgage foreclosure help, mortgage foreclosure, mortgage help. Since Google will not show any result whose frequency is under a specific threshold, only 20 cities' data is presented. Considering the validity and feasibility of both residential energy efficiency & mortgage default risk, we set the time range to the most recent year on Google Trends.

In data preprocessing we find the mean of each term and then normalize it by using the following formula:

$$X(i)= \frac{X(i)}{\max(X(i))} \times 100$$

where X(i) the i^{th} instance in vector X.

Figure 2 depicts an example of the Mortgage Default Risk Index (MDRI) in New York City. This plot shows the normalized MDRI of New York from 03/20/2016 to 04/19/2017. It combines terms of foreclosure, foreclosures, foreclosure help, housing assistant, mortgage assistant, mortgage foreclosure, and mortgage help. As we can see, after 07/03/2016 the default frequency experienced a sharp increase. Then after a slight drop, the frequency continued to increase. It reached its peak on 08/07/2016. Mortgage default rates have been struggling to rise since the Brexit vote (June 23, 2016). It is reasonable to see this fluctuation. After the result of Donald Trump's victory in 2016, the mortgage market shows another big fluctuation, and the graph also depicts this pattern.

We then calculate the mean of each city's data to generate the MDRI for each city.

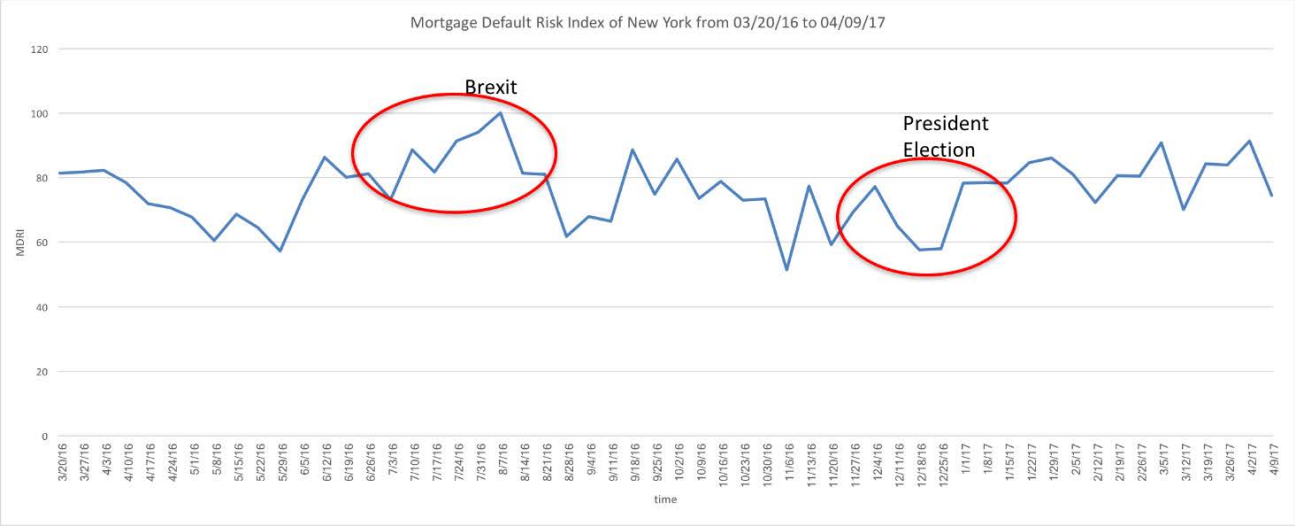


Figure 2: Mortgage default risk index of NY from 03/20/16-04/09/17

According to the above description, we form two datasets with 18 instances and 8 attributes

each, depicted in the following tables 2 & 3.

CITIES	MDRI	Ener_Mean	HOTWATER	HEATING	BASELOAD	COOLING	HEATING_FUEL_TYPE	TOTAL_COST
PHOENIX	67.9595086	180	106.62	2247.93	4634.72	4782.08	0	1683.47
SAN DIEGO	53.4210526	132	141.74	6018.06	5126.26	0.00	0	2088.46
DENVER	62.4231951	168	202.07	25490.38	4857.55	660.84	0	3869.09
TAMPA	64.3068888	132	110.76	1784.79	5412.40	6920.31	0	1843.90
ORLANDO	67.0401966	120	118.90	1544.34	5028.53	4240.20	0	1476.18
ATLANTA	60.7356715	132	196.17	10236.73	6210.31	2277.36	0	2484.60
CHICAGO	70.2938072	180	199.17	1607.60	5446.32	1335.65	1	2428.05
BOSTON	58.8041126	144	200.27	1270.32	6914.46	622.66	1	4096.10
DETROIT	70.6743567	168	208.70	1287.45	5394.10	757.40	1	2199.45
NEW YORK	76.308309	180	176.94	1850.16	5068.15	1544.92	2	5232.20
CHARLOTTE	63.7825572	144	174.52	15738.03	6075.97	2520.22	0	3026.26
PHILADELPHIA	66.7232691	180	193.90	1518.23	5516.94	2173.77	1	2921.26
SAN ANTONIO	73.2672518	144	141.63	5092.40	5430.33	5402.17	0	2092.79
DALLAS	51.8582715	156	150.56	6621.80	5095.35	4062.98	0	2085.62
HOUSTON	61.8290669	144	141.69	5536.53	5662.72	6744.86	0	2337.39
SEATTLE	71.4712578	144	200.77	664.96	5706.96	79.28	1	1462.37
AUSTIN	54.7945205	180	122.15	7063.87	4932.92	7837.77	0	2543.65
LOS ANGELES	76.2028127	132	161.58	1992.22	5750.51	101.52	0	1537.58

Table 2: Whole dataset calculated by Mode

CITIES	MDRI	Ener_Mode	HOTWATER	HEATING	BASELOAD	COOLING	HEATING_FUEL_TYPE	TOTAL_COST
PHOENIX	67.9595086	180	106.62	2247.93	4634.72	4782.08	0	1683.47
SAN DIEGO	53.4210526	120	141.74	6018.06	5126.26	0.00	0	2088.46
DENVER	62.4231951	180	202.07	25490.38	4857.55	660.84	0	3869.09
TAMPA	64.3068888	108	110.76	1784.79	5412.40	6920.31	0	1843.90
ORLANDO	67.0401966	108	118.90	1544.34	5028.53	4240.20	0	1476.18
ATLANTA	60.7356715	120	196.17	10236.73	6210.31	2277.36	0	2484.60
CHICAGO	70.2938072	180	199.17	1607.60	5446.32	1335.65	1	2428.05
BOSTON	58.8041126	120	200.27	1270.32	6914.46	622.66	1	4096.10
DETROIT	70.6743567	180	208.70	1287.45	5394.10	757.40	1	2199.45
NEW YORK	76.308309	180	176.94	1850.16	5068.15	1544.92	2	5232.20
CHARLOTTE	63.7825572	120	174.52	15738.03	6075.97	2520.22	0	3026.26
PHILADELPHIA	66.7232691	180	193.90	1518.23	5516.94	2173.77	1	2921.26
SAN ANTONIO	73.2672518	120	141.63	5092.40	5430.33	5402.17	0	2092.79
DALLAS	51.8582715	120	150.56	6621.80	5095.35	4062.98	0	2085.62
HOUSTON	61.8290669	120	141.69	5536.53	5662.72	6744.86	0	2337.39
SEATTLE	71.4712578	120	200.77	664.96	5706.96	79.28	1	1462.37
AUSTIN	54.7945205	180	122.15	7063.87	4932.92	7837.77	0	2543.65
LOS ANGELES	76.2028127	120	161.58	1992.22	5750.51	101.52	0	1537.58

Table 3: Whole dataset calculated by Mean

4.3 Implementation

We use MATLAB and Python to build our models, and their summary is depicted below. Python, by definition, is a programming language. Python also consists of an extensive standard library. This library is aimed at programming in general and contains modules for OS specific stuff, threading, databases and statistical modeling, etc. MATLAB is a commercial numerical computing environment and programming language [6].

4.3.1 Correlation Analysis

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two numerically measured, continuous variables. This particular type of analysis is useful when the goal is to establish connections between variables. If significant correlation is found between two variables it means that when there is a systematic change in one variable, there is also a systematic change in the other; the variables alter together over a certain period of time. If there is correlation found, depending upon the numerical values measured, this can be either positive or negative.

Positive correlation exists if one variable increases simultaneously with the other, i.e. the high numerical values of one variable relate to the high numerical values of the other. Negative correlation exists if one variable decreases when the other increases, i.e. the high numerical values of one variable relate to the low numerical values of the other [7].

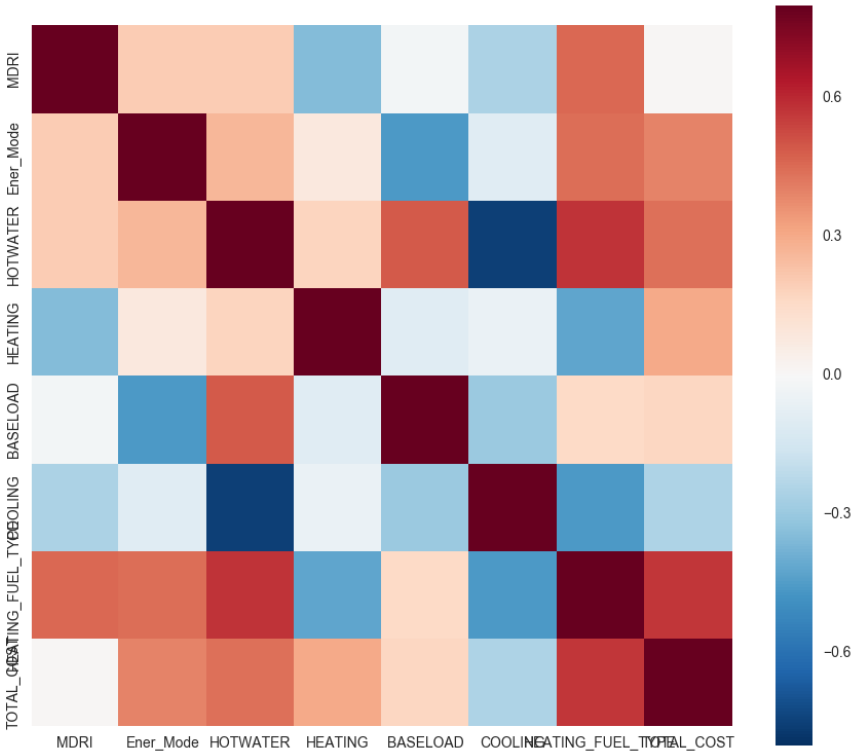


Figure 3: Covariance Matrix of Eight Attributes Including *Ener_Mode*

In our experiments, we use a correlation matrix to detect the relationship between the attributes. We performed a correlation analysis on both datasets. The covariance matrix of Ener_Mode is depicted in Figure 3. The results show that the correlation between *MDRI* and

Ener_Mode is higher than between MDRI and *Ener_Mean*. Hence in the following experiments we decide to use the dataset with *Ener_Mode* rather than *Ener_Mean.* In the above graph we plot a heatmap representing the covariance matrix of 8 attributes. As the color bar shows, the deeper the color, the higher the correlation. We can see that *HEATING_FUEL_TYPE* has the highest correlation with *MDRI*, whereas *BASELOAD* has zero correlation. The correlation between *Ener_Mode* and *MDRI* is about 0.25. It is relatively higher than the correlation between baseload/total cost and MDRI, from which we can see a promising relationship between home energy efficiency and mortgage default risk.

4.3.2 Multi-Linear Regression

In statistics linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or predictor variables) denoted X. We choose *Ener_Mode, HOTWATER, HEATING, BASELOAD, COOLING, HEATING_FUEL_TYPE* and *TOTAL_COST* as our X and *MDRI* as our Y to perform the multiple linear regression.

As the table 4 shows below the R-squared value is promising, though not ideal, as it indicated that the model accounts of about 34.8% of variation. In a linear regression model, a small p-value (generally less than 0.05) indicates a strong evidence to accept our model. Additionally, of the seven predictor variables used, the *Ener_Mode* has the highest p-value (0.69), whereas *HEATING_FUEL_TYPE* has the lowest p-value (0.25), both of which are too high when compared to standard thresholds. Noting that our set of predictor variables does not include macroeconomic factors that should affect mortgage default in a first-order sense, such as changes in interest rates or levels of unemployment, we find these results suggest that there is likely an effect of energy efficiency on mortgage default. The type of heating in a house has the most influence on mortgage default risk. By the correlation analysis above, *Ener_Mode* has high multicollinearity with *HEATING_FUEL_TYPE*, which can explain the poor performance of *Ener_Mode*. We then look at its comparative VIF and Cp.

OLS Regression Results

Dep. Variable:	MDRI	R-squared:	0.348
Model:	OLS	Adj. R-squared:	-0.108
Method:	Least Squares	F-statistic:	0.7634
Date:	Sun, 30 Apr 2017	Prob (F-statistic):	0.630
Time:	14:46:27	Log-Likelihood:	-57.149
No. Observations:	18	AIC:	130.3
Df Residuals:	10	BIC:	137.4
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	64.5062	35.744	1.805	0.101	-15.137	144.150
Ener_Mode	0.0457	0.111	0.409	0.691	-0.203	0.294
HOTWATER	-0.1525	0.213	-0.715	0.491	-0.628	0.323
HEATING	0.0008	0.001	0.702	0.499	-0.002	0.003
BASELOAD	0.0048	0.009	0.527	0.610	-0.015	0.025
COOLING	-0.0006	0.001	-0.471	0.648	-0.003	0.002
HEATING_FUEL_TYPE	17.2054	14.119	1.219	0.251	-14.255	48.665
TOTAL_COST	-0.0064	0.006	-1.104	0.296	-0.019	0.007

Omnibus:	0.678	Durbin-Watson:	2.533
Prob(Omnibus):	0.712	Jarque-Bera (JB):	0.165
Skew:	0.234	Prob(JB):	0.921
Kurtosis:	3.012	Cond. No.	1.90e+05

Table 4: Summary of OLS Regression Results

The variance inflation factor (VIF) of a parameter variable tells us how much the squared standard error of that variable is increased by having the other predictor variables in the model. In other words, it measures the severity of collinearity attributed to the predictor variable that is left out of the model. The higher the VIF, the worse. The VIF of each of the seven predictor variables is shown in Table 5.

Predictor Variable	VIF
Ener_Mode	1.042
HOTWATER	1.043
HEATING	1.137
BASELOAD	1.000
COOLING	1.068
HEATING_FUEL_TYPE	1.264
TOTAL_COST	1.000

Table 5: VIF of Each Predictor Variable

The predictor variable that has a relatively high VIF is the heating type of the house, though all the seven variables give small values. This result is also consistent with the correlation analysis.

We then remove *HEATING_FUEL_TYPE* for the flowing experiments.

The centered predictor values (Cp) - which measures a model's capability of predicting new responses for MDRI – was calculated for every combination of the six predictor variables. In the table below we mark variables that are included in the model with 'X'. In our experiments, we try to remove each attribute each time to see the value of Cp. We see that every value of Cp is greater than (p+1), where p is the number of predictor variables used. This may mean that the model is overfitting. As we can see from the table 6, all the values are larger than (p+1). The model with the highest Cp omits *HEATING* as a predictor variable, which is the poorest predictor. Whereas the model using the set {*HOTWATER, HEATING, BASELOAD, COOLING, TOTAL_COST*} has the lowest

Ener_Mode	HOTWATER	HEATING	BASELOAD	COOLING	TOTAL_COST	Cp
×	×	×	×	×		13.524
×	×	×	×		×	13.685
×	×	×		×	×	13.820
×	×		×	×	×	17.035
×		×	×	×	×	13.850
	×	×	×	×	×	13.551

Table 6: Cp of Each Predictor Variable

Cp and therefore would be the optimal choice, but we will not consider it because we do not want to remove *Ener_Mode* in our experiments. Overfitting in our result seems reasonable because we have only 18 instances to be trained in the model. The model would fit better if a larger size sample is used.

We then detect the outliers in dataset. Here we use Cook's Distance. Cook's Distance is used to indicate influential data points that are particularly worth checking for validity. Large values of Cook's distance can be seen as an outlier in the entire dataset.

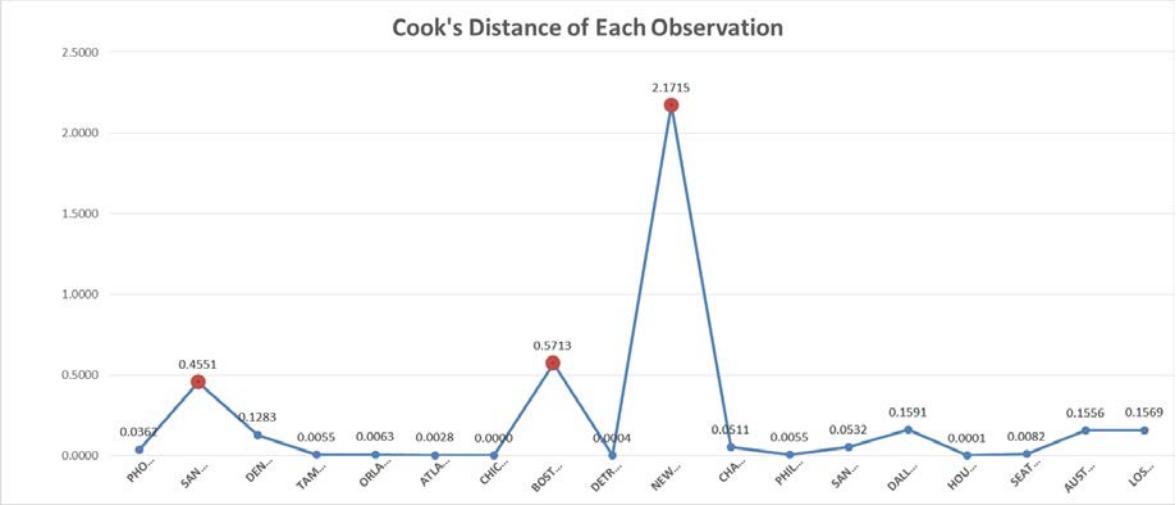


Figure 4: Cook's Distance for Each City

The Cook's Distance for each observation is depicted in figure 4. The mathematical outliers are highlighted in red. Essentially, Cook's distance combines residuals and leverage in a single measure of influence. Because of the lack of instances, we only remove the point with the highest Cook's Distance, which is New York. The updated dataset produces a higher R-square of 0.412.

As demonstrated above, linear regression with R-square of 0.348 does not have a high performance because of the quality of data is not able to explain the relationships in a defined way. Therefore, we can infer that these variables do not have an obvious linear relationship. We use a non-linear regression model (regression tree) to improve our results.

4.3.3 Regression Tree

A regression tree is built through a process known as binary recursive partitioning, which is an iterative process that splits the data into partitions or branches, and then continues splitting each partition into smaller groups as the method moves up each branch.

We trained the regression tree on both datasets. A pruned regression tree trained on *Ener_Mode* is shown in Figure 5. As we can see the tree starts from the *HEATING* attribute, which means that heating presents a significant effect in mortgage default risk. There is no *Ener_Mode* in this tree.

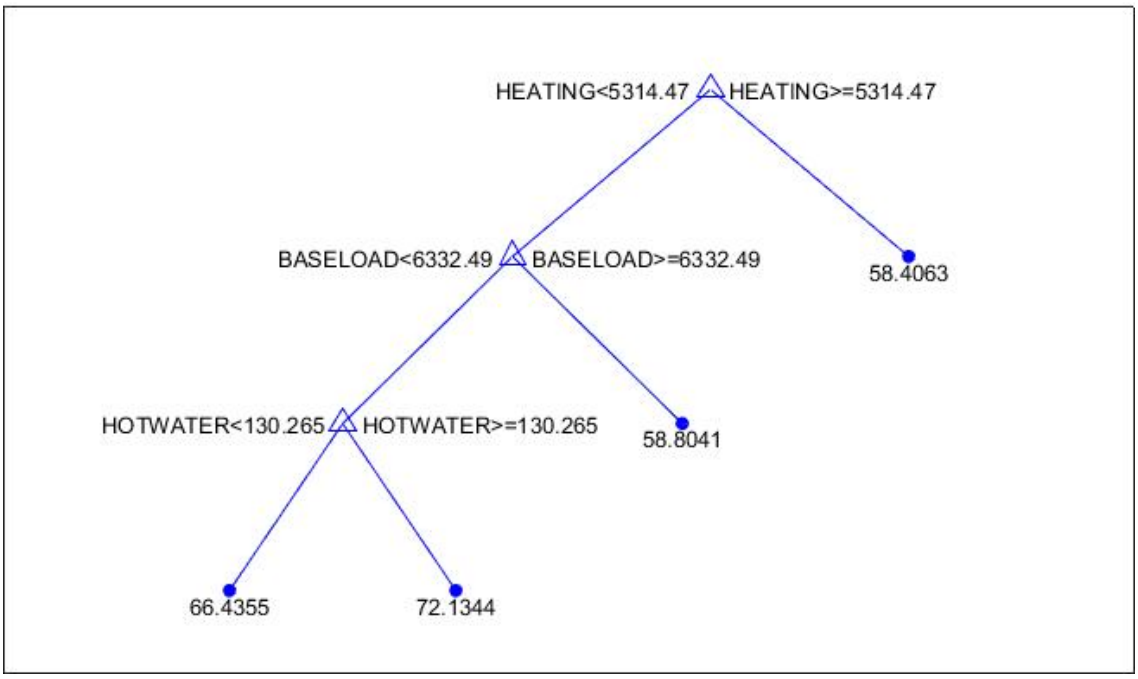


Figure 5: Pruned Regression Tree on *Ener_Mode* dataset

To see whether home energy efficiency has an important effect in the regression model, we try to train the model with *Ener_Mode* and without *Ener_Mode*. If the performance is worse after removing the attribute, we can say that *Ener_Mode* has an important effect in predicting the MDRI to some extent. Since we do not have a test data set, the method to calculate the prediction error is called the in-the-sample test. Specifically, we use cross-validation. This method calculates the Mean-Square-Error (MSE) by dividing the entire dataset into k groups without replacement. Use k-1 groups as the training set and the remaining one as test set. Considering the numbers of our instances, we set k=8.

MATLAB gives us a MSE of 42.22 on the initial dataset. After removing *Ener_Mode*, the MSE grows to 50.11. The increasing of MSE indicates that if the key variable *Ener_Mode* is ignored, our model performance will be negatively influenced, which implies that home energy efficiency plays a very important role in building the regression tree model.

5. BUSINESS DRIVERS

We have identified business drivers for continued success and sustainability of EnerScore and the proposed Portfolio Risk Assessment for Mortgages. Identifying EnerScore's business drivers should enhance EnerScore's capacity to implement this new business tool. We are not considering factors that EnerScore cannot influence or use to its benefit at this stage, such as economic conditions or trade relations with other nations. Monitoring these business drivers will help EnerScore achieve technological innovation and superior marketing of its product.

- **Environmental:** This aims at recognizing the responsibility that each of us have to ensure that we protect our natural resources for future generations. Through this project, we are taking steps to ensure energy is used efficiently and more people are aware about energy efficiency.
- **Financial:** EnerScore built an API that is designed for real estate sites, but the rate of adoption was low. Through this project, EnerScore ventures into a different market domain that targets financial industry end users, thereby increasing sales and market potential.
- **Market:** Create the EnerScore brand in the market for innovation and reputation

6. SYSTEM REQUEST

- **Project Sponsor**: EnerScore (Tom Witkin, Brian Butler, and Kenn Butler)
- **Business Needs:**
  - Increase rate of adoption in business
  - Increase sales
  - Increasing market share through innovation
  - Create industry standard, brand
- **Business Requirement:**
  - Develop the Portfolio Risk Assessment for Mortgages model to API to integrate with EnerScore website.
  - System should be able to define a return for mortgage originators.
  - System should be able to give financial investment firms the ability to manage the risk of portfolios of securitized mortgages.
  - Deliver a model that establishes a risk reward ratio for a portfolio of mortgages based solely on energy efficiency.
  - System should utilize energy efficiency, but not underlying finances.
  - Underlying model created should not target certain demographics or particular community; focused groups.
  - Should be developed in a way that is customizable in future for varied customers.

  - Create a marketing strategy for target audience (e.g., mortgage originators, retail investors, community banks, credit unions, or investment banks).
- **Potential Benefits:**
  - Promote Energy Efficiency (impact on environment/society).
  - Promote EnerScore's brand.
  - Enable a faster return on investment.
  - Improve decision-making process for community banks, credit unions, loan originators, investment banks and other investors.
  - Identify new clients and support existing client base better.
- **Special Issues and Constraints:**
  - From a marketing perspective, this as a strategic system. The ability to offer a risk reward ratio for a portfolio of mortgages is critical in order to remain competitive in market.
  - Project timeline constraint – 4 months.
  - Lack of access to default data and lack of funds

7. TECHNICAL FEASIBILITY ANALYSIS

- **Current System ("As Is")**
  - EnerScore has a database of individual records such as public record and tax record data.
  - The database is MongoDB and is used to query data. The data is comprised of two sets: one with 75 million estimated scores, and the other with ½ million HERS scores resulting from on-site human inspections.
  - The current EnerScore model in place is built by JavaScript and focuses on HERS rating at its core is derived from various sets of public data and records. The Home Energy Rating System (HERS) Index is an industry standard by which a home's energy efficiency is measured. It is also the nationally recognized system for inspecting, testing and calculating a home's energy performance. The outputs of the current EnerScore model are the efficiency levels from A to F (a result of partitioning the 200-point range of positive values into subintervals of size 12) where A represents net zero energy use.
  - The data queried from Mongo DB is processed online by Java script.

- **Recommended System ("To Be")**

  The recommended system is based on the varied audience targeted by EnerScore, number of users per year, and costs involved. Integrate the EnerScore Portfolio Risk Assessment for

Mortgages model to API for the EnerScore website. This solution fits the criteria mentioned above.

Additionally, this website should fulfill the following requirements for ensuring user satisfaction and usability:

- Developed with basic functionality, fit to be disability friendly
- Developed with easy to read font styles
- Developed for good visual appeal
- Developed with consistency of design, clarity of data and simplicity in functionality

The users accessing this decision aid will also be provided with a accessibility guide with the functionalities and navigation

Non-functional requirements of the recommended system are:

- **Operational:** The system can run on all web browsers.
- **Performance:** The system should have maximum response time of ten seconds**.** It should be able to support minimum 200 users and scalable up to 300 users for a start**.** The system should be available 24 hours.
- **Security:** Safety of personal information entered should be secure.
- **Cultural:** Display information on the front-end in English and understandable format**.** Necessary to present the information in an organized way and cater to different audience and sensibilities.
- **Recovery:** If a database crashes then the uptime would be 99%. The downtime estimated would be Two Nines equals to 7 hours and 12 minutes downtime in 30 days.

After assessing the risk according to the different factors, we can confirm that the prototype can be developed. Please see Table 9 below.

	Project team	End user
	Risk Associated	Risk associated
Familiarity with application	Low	Medium
Familiarity with Technology	High	Low
Project Size	Medium	Low
Compatibility	High	Low

Table 9: Factors that affirm that the prototype can be developed

8. PROJECT METHODOLOGY

We recommend Rapid Application Development [8] as a project methodology. Rapid application development is a methodology that uses minimal planning in favor of rapid prototyping. A prototype is a working model that is functionally equivalent to a component of the product.

In the RAD model, the functional modules are developed in parallel as prototypes and are integrated to make the complete product for faster product delivery. Since there is no detailed preplanning, it makes it easier to incorporate the changes within the development process.

RAD projects follow iterative and incremental model and have small teams comprising of developers, domain experts, customer representatives and other IT resources working progressively on their component or prototype. The phases of RAD covered extensively for this project are, Business Modelling and Data Modelling. The advantages of the RAD Model are as follows:

- Changing requirements can be accommodated.
- Progress can be measured.
- Productivity with fewer people in a short time.
- Reduced development time.
- Quick initial reviews occur.
- Encourages customer feedback.

This methodology is performed iteratively. This cyclic process of analyzing and refining a process is intended because it will ultimately help improve the quality of understanding and delivered product. Since the product can be modularized in 2-3 months, and the resources have thorough business knowledge from research, and highly skilled developers. In addition, we will be maintaining documentation throughout the system development life cycle, which makes this an apt project methodology model.

9. STAFF PLAN

In order to meet the client requirements effectively as a team, we designed standard roles and responsibilities for our team.

- **Market Analyst**
    - Responsible for understanding the current market trends, researching why this project adds value to the current market, the organization and how the developed product can be effectively commercialized for target audience.
- **Data Scientists**

- Responsible for understanding client requirements, analyzing data and developing quantitative model as requested by client.
- Since the timeline is a constraint for the development of this model, two members will be working in this role.

- **Business Analyst:**

  The responsibilities include:
  - Interviewing stakeholders, coordinating with client, handling each phase of the project acting as a liaison between developers and clients.
  - Understanding current system technically and how new system can integrate with their current system by performing gap analysis.
  - Maintaining necessary documentation.

10. ANALYSIS STRATEGY

Analysis strategy is based on some important factors like depth of information and integration of information.

The depth of information includes, obtaining not only facts and opinions of the client and advisors, but also gain a thorough understanding of why those facts and opinions exist through research and how they could be incorporated while developing the project.

For example, interviews and joint application sessions are very useful at providing a good depth of rich and detailed information and helping the analyst to understand the project goals and how to develop the to-be system. Document analysis and observation are useful for obtaining facts, but provide a very generic outlook towards the project. Questionnaires can provide a medium depth of information, soliciting both facts and opinions and enable the analyst to understand the current system.

For this reason, we will be conducting document analysis to give us facts about the EPRAM project, followed by interview sessions with appropriate questionnaires.

Integration of information refers to gathering information from various sources. For example, this means referring to different research papers for mortgage backed securities, contacting CoreLogic, Bloomberg for foreclosure data, speaking to client for additional information on their current system. Referring to many documents and people might provide us some information that might conflict with our understanding of the system, research conducted or facts discussed with the client.

Combining this information and attempting to resolve differences in opinions or facts, explaining the discrepancy and attempting to refine the information is a vital step in integrating the information.

Thus, after we gathered information from different sources we documented our understanding and next steps for the client by creating a business requirement document.

Below are the details of the various requirement-gathering strategies used.

Document analysis:

The documents referred for document analysis included, IMT UNC Home EE Mortgage Risks, residential mortgage default risk and the loan-to-value ratio, mortgage default risk new evidence from internet search queries. These documents provided us the necessary information about correlation that exists between default risk and energy efficiency, mortgage default risk and how to achieve a viable quantitative model.

Client communications:

The team kept in touch with the client contact – Tom Witkin -- with understanding of the project and issues faced, and developments achieved.

Interviews:

Conducting an interview is a low cost technique that has the maximum client involvement and is one of the most effective requirement gathering methods. Hence, interviews with client were conducted to ensure the team understands the business problem, current system model, and what expectations of the to-be system from the client. Appropriate questions were designed for these interview sessions to ensure that the team gains a thorough understanding of goals within an allocated time.

Questions used for requirement gathering were categorized into feasibility analysis of current and to-be system, risk assessment, economic feasibility and analysis of system. Some of the sample questions were:

- The benefits that the EnerScore Portfolio Risk Assessment for Mortgages (EPRAM) would create for the company
- Any special issues or constraints relevant to implementation of system? Example (security clearance needed to work on certain data, timeline constraints)
- Discuss current system and to - be system.
- What type of data and volume are we looking at? Structured or Unstructured?

- Who will use the system – understand customer segment to design the perfect end application
- What is your current business model and service model? How do you make profit from the EnerScore service?
- Are borrowers' incomes, market conditions considered in the data set provided?
- Who would serve as the primary point of contact on business side?

Based on requirement gathering of data and information, constraints and risks are highlighted as follows:

Constraints

- Ensure the model created addresses the concerns surrounding ethics issue
- Project timeline constraint – 4 months
- Lack of access to default data
- Budget constraints

Risks

- Lack of accuracy if foreclosure data is considered instead of default data
- No analysis conducted to ensure that modelling this solution will satisfy EnerScore's business needs
- No evaluation, testing plan in place for the output of the To – be system model

11. MARKET ANALYSIS

11.1. Product Overview – EPRAM

The EPRAM built and developed by EnerScore and WPI collaboratively, is aiming at providing the strong correlation between the energy efficiency data and default risk data.

EnerScore developed the prototype of EPRAM based on the HERS scoring data and WPI utilized the regression and linear regression analysis on city-level foreclosure data to further prove the correlation between the energy efficiency data and the default risk data. The development of the correlation between the city-level foreclosure data would be the first step for the EnerScore risk assessment model. In the later stage the analysis on individual-level foreclosure data would be performed to further prove the correlation and improve the prediction accuracy of the risk assessment model. At current stage, the EPRAM could be considered as the first generation model.

The ultimate purpose of developing EPRAM is to provide loan originators or mortgage-backed securities underwriters a tool to evaluate the mortgage default risk more comprehensively based on available energy efficiency data and mitigate the risk and probability of possible subprime crises occurrences in the future.

11.2. Industry Background and Outlook for MBS Market

The correlation analysis between the subprime mortgage crisis and the MBS market

The mortgage market is the largest long-term debt market in the US. Based on data released by Federal Reserve, as of the fourth quarter of 2016, the mortgage debt outstanding in the United States totaled 14.29 Trillion US Dollars, among which the mortgages for one-to-four family residences account for approximately 72% of the total market debt outstanding [9]. From the above data it is evident that the mortgage played a critical role in the housing market when Americans chose to purchase and finance their homes.

One of the reasons for the development of Mortgage-Backed Securities is for liquidating the illiquid and inefficient mortgages. At the primary market the mortgage originators underwrite the home mortgages to borrowers. It was not until 1968 that the secondary market was established and developed for the residential mortgages to trade as mortgage-backed securities. After the secondary market for the loan mortgage was established, mortgage originators do not have to hold the loans they originated until the last payment date but could instead choose to trade the loan to investors as mortgage-backed securities. [10

The financial crisis of 2008 is also known as the subprime mortgage crisis, the biggest financial recession since the Great Depression in the 1930s. Consequently, the exuberant housing bubble in the United States before 2008 could be blamed as the major trigger for the financial crisis worldwide. How could the housing bubble in the United Stated drag down the economy on a worldwide scale? A subprime loan is the situation where the mortgage originator, or the lender, lends the funds to people with poor credit history and high risk of default.

Figure 6 shows the subprime mortgage originations from 1996 to 2008. It is shown that in 2006 $600 Billion of subprime loans were originated, most of which were securitized, and in that year subprime lending accounted for 23.5% of all mortgage originations. Although the mortgage originators who originated the subprime loans to the substandard borrowers should be considered the biggest culprit, the triggers resulting in the 2008 financial crisis are a combination of several factors that are far from being simple. The subprime mortgage originators, the over-optimistic homebuyers, the investment banks, the bond rating agencies, the investors, such as hedge funds willing to invest in MBS, and the Federal Reserve all share some blame in triggering the outbreak of the subprime mortgage crisis [11].

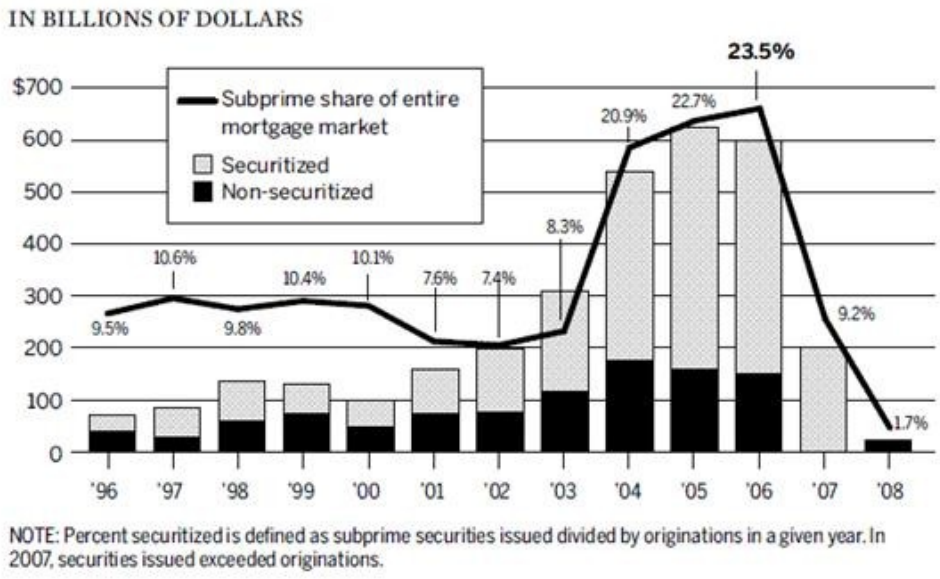


Figure 6: Subprime Mortgage Originations [12]

The methodology flaw of the risk assessment model related to the subprime mortgage crisis

As mentioned in the book *The Subprime Solution: How Today's Global Financial Crisis Happened, and What to Do about it* by Robert J Shiller [13], the perception that real estate prices only go up, year after year, established an atmosphere that invited lenders and financial institutions to loosen their standards and risk default. Before the outbreak of the subprime mortgage crisis in 2008, the housing market and the MBS market were like a carnival that had attracted millions of participants such as mortgage lenders, investment banks, and various investors. Every participant was anxious to take a share of the feast and completely ignored the looming housing bubble risk and the potential crisis. In order to stimulate the economy and suppress the recession that resulted from the dotcom bubble in 2000 and the September 11th terrorist attacks, the Federal Reserve lowered interest rates substantially from 6.5% in May 2000 to 1.75% in December 2001 and then all the way down to 1% in June 2003 to create capital liquidity in the market.

With the abundance of capital with low-interest rates, mortgage originators began to loosen their credit standards and even lend funds to people with flawed credit history for more profitability. Then the banks sold the mortgages originated to the investment banks, who then securitized the all the mortgages into bonds and then sold them to investors.

Since the mortgage originators could easily sell the mortgages to investment banks for securitization and then free up more capital for lending, the mortgage originators were more

than willing to originate more mortgages to even more substandard borrowers. The more mortgages the mortgage originators lend, the more origination fees and profits they could earn. The more mortgages the mortgage originators sell to the investment banks, the more profit the investment banks could earn by repackaging the mortgage pools into bonds and selling them to investors.

The eruption of the subprime crisis could not be boiled down to only one reason. The methodology and system flaw in the existing risk assessment model could be one of the reasons for the subprime crisis. After the outbreak of the subprime mortgage crisis, the financial institutions have implemented more stringent underwriting standards, and the investors have been more cautious than ever before. Moreover, the United States also issued new regulations to address the effects of the subprime mortgage crisis. The Dodd–Frank Wall Street Reform and Consumer Protection Act was signed and issued in July 2010 to address all the negative effects of the subprime mortgage crisis.

11.3. Targeted Customer and Market Analysis

Mortgage originators (small-sized loan originator)

Because of the reputation and brand awareness of the large banks or financial institutions, the residential loans originated by large banks could be securitized and sold to investors in a relatively easy way for the purpose of liquidating the originated mortgages. However, the loans originated by small-sized banks would be more difficult to be sold to investors as MBS. The EPRAM risk assessment model could add some edges for increasing the MBS underwriters and investors' confidence towards the residential loans originated by small-sized institutions.

Mortgage-backed security underwriters

The Utilization of the EPRAM could provide the mortgage-backed security underwriters an evaluation tool that could be linked with relatively high-quality residential loans. This could also be considered as a useful tool for more accurate MBS pricing.

The market opportunities analysis for the EPRAM

The collapse of the mortgage market and the MBS market after the outbreak of the subprime crisis led and directed the overoptimistic and exuberant market to return to a normal and circumspect market again. The failure of the market made investors more wary of the potential risks. New regulations from the governments, the more stringent underwriting policies by financial institutions, and the more conservative actions of various investors all indicate that new technology innovations for hedging risks will be necessary for the capital market. This also indicates new opportunities for technologies and market product innovations in the financial markets.

EPRAM could become a risk assessment tool that enhances risk assessment in the residential mortgage and MBS mortgage markets. The implications of the application of EPRAM could be profound. A new credit review and risk assessment model would make the mortgage underwriting process more objective, overall and transparent. The implication of the risk assessment model would also help to rebuild a sound financial market and system and restore the investors' confidence in the whole financial market.

The new risk assessment modeling for the mortgage originators

In the subprime mortgage crisis the mortgage originators are blamed and take one of the major responsibilities for loosening the credit standards for substandard borrowers. After the crisis, the loan originators became unwilling to take any risks to accept substandard borrowers, which means that there would be more difficulties and a smaller chance of getting residential mortgages for borrowers. Therefore, there is a necessity for the mortgage originators to have a more overall and objective methodology in their risk assessment models.

The strengthened supervision of the MBS market

The strengthened supervision of the MBS market after the subprime crisis would make it a necessity and priority for investment banks or the securities dealers to choose and sell the mortgages originated and underwritten using a more accurate and objective credit review and default risk assessment model.

The discretion of the investors after the subprime mortgage crises

After surviving the subprime crisis, the various investors would become more risk averse in investing in the MBS market. A new and overall credit review and default risk assessment system would make the underwriting process more reliable and could earn the loan originators and the investment banks a competitive edge in selling the mortgage-backed securities to the investors.

Government regulation related to the strengthening MBS market

With the government issuing new regulation of supervising the MBS market in a more stringent way, The MBS market needs more accurate and objective risk assessment models to rebuild the investors' confidence and the market's reputation.

12. FUTURE RESEARCH AND NEXT STEPS

Considering there are many other factors influencing energy efficiency, such as interest rates, employment levels, and FICO scores, it is reasonable to have an R-square value of 0.348 in the regression model. But due to the limited data, it is difficult to establish a definite existence of a linear relationship between EnerScore and default risk.

In future analysis, we would like to establish a stronger prediction model, given additional information from EnerScore. The first step would be to perform regressions using standard predictors of default – macroeconomic factors given city-level data (e.g., unemployment level, change in unemployment level, yield on the 10-year bond, and change in the yield on the 10-year bond). Then we would suggest completing the new regressions with EnerScore data included.

Based on the updated R-square/adjusted R-square, we would hope to establish whether or not the EnerScore data provides more predictive power to the regression. Then by calculating VIF or Cp, we can check the collinearity between EnerScore data and other predictor variables. Further, leverage and residuals will be considered to detect outliers.

That being said, a larger dataset is necessary to achieve the precision standards. Access to more city-level data is necessary to build the model to further demonstrate these relationships. Additionally, access to individual MDRI for each address, which can be mapped in the EnerScore dataset, will further boost the credibility of the model.

From our research, CoreLogic provides a good platform for us to access the mortgage default risk data. Diving in deeper to understand how CoreLogic data can be used for the model is essential.

In addition, more efficient statistical models like clustering analysis and support vector machines would be helpful for this project. These methods can help explore the relationship between MDRI and home energy efficiency more thoroughly and clearly.

REFERENCES:

[1]. Sharpe, William F. "The Sharpe Ratio." The Journal of Portfolio Management 21, no. 1 (1994): 49-58. doi:10.3905/jpm.1994.409501.

[2]. Chauvet, Marcelle, and Stuart A. Gabriel. "Mortgage Default Risk: New Evidence from Internet Search Queries." SSRN Electronic Journal. doi:10.2139/ssrn.2816207.

[3]. "Mortgage Default Management Solutions." Mortgage Default Management. Accessed May 04, 2017. http://www.corelogic.com/solutions/mortgage-default-management-solutions.aspx.

[4]. Ruppert, David. Statistics and data analysis for financial engineering. New York: Springer, 2011.

[5]. Introduction to Data Mining P.-N. Tan, M. Steinbach, V. Kumar. Addison-Wesley 2005. ISBN-10: 0321321367 ISBN-13: 9780321321367

[6]. "Python vs Matlab." Python vs Matlab — Pyzo - Python to the people. Accessed May 04, 2017. http://www.pyzo.org/python_vs_matlab.html.

[7]. Correlation Analysis - Market Research. Accessed May 04, 2017. http://www.djsresearch.co.uk/glossary/item/correlation-analysis-market-research.

[8]. Foundations of Software Testing: ISTQB Certification

[9]. "Mortgage Debt Outstanding." FRB: Mortgage Debt Outstanding, March 2017. Accessed May 04, 2017. https://www.federalreserve.gov/econresdata/releases/mortoutstand/current.htm.

[10]. Financial Institutions, Markets and Money, 12th Edition, David. S. Kidwell/David W.

Blackwell/David A.Widbee/Richard W. Sias

[11]. Petroff, Eric. "Who Is To Blame For The Subprime Crisis?" Investopedia. March 14, 2017. Accessed May 04, 2017. http://www.investopedia.com/articles/07/subprime-blame.asp.

[12]. Source: Inside Mortgage Finance, Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States, National Commission on the Causes of the Financial and Economic Crisis in the United States

[13]. Shiller, Robert J. The subprime solution: how today's global financial crisis happened and what to do about it. Princeton, NJ: Princeton University Press, 2012.

APPENDIX – LINEAR REGRESSION:

```
import pandas as pd
import numpy as np
import csv
from numpy import matrix
from numpy.linalg import inv
from statsmodels.formula.api import ols
#from pandas.stats.api import ols
import statsmodels.api as sm
from __future__ import print_function
from statsmodels.compat import lzip
import statsmodels
import statsmodels.formula.api as smf
import statsmodels.stats.api as sms
import matplotlib.pyplot as plt
from statsmodels.stats.outliers_influence import OLSInfluence
from __future__ import print_function
import itertools

data = pd.read_excel('dataset.xlsx',sheetname = 'mode',parse_cols = "B:I")
# data = raw_data.parse('mean')
# data = pd.DataFrame(raw_data)
size = data.shape
data.head()

# correlation heatmap for nominal sttributes
```

```
import seaborn as sns
sns.set(style="white")

# Load the datset of correlations between cortical brain networks
corrmat = data.corr()
# print(corrmat)
# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(11, 9))
cmap = sns.diverging_palette(220, 10, as_cmap=True)

# Draw the heatmap using seaborn
sns.heatmap(corrmat, vmax=.8, square=True)
plt.savefig('heatmap_mean')
# plt.show()

# correlation heatmap for nominal sttributes
import seaborn as sns
sns.set(style="white")

# Load the datset of correlations between cortical brain networks
corrmat = data.corr()
# print(corrmat)
# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(11, 9))
cmap = sns.diverging_palette(220, 10, as_cmap=True)

# Draw the heatmap using seaborn
```

```
sns.heatmap(corrmat, vmax=.8, square=True)

plt.savefig('heatmap_mode')

# plt.show()


model =
ols('MDRI~Ener_Mode+HOTWATER+HEATING+BASELOAD+COOLING+TOTAL_COST',data).fit()

print(model.summary())


r = model.rsquared

sigma_square = model.ssr/(size[0]-1)


factors =
('Ener_Mode','HOTWATER','HEATING','BASELOAD','COOLING','HEATING_FUEL_TYPE','TOTAL_C
OST')


for item in itertools.combinations(factors,6):

   P = '+'

   L = '~'

   name_6i = ('MDRI',L,item[0],P,item[1],P,item[2],P,item[3],P,item[4],P,item[5])

   name_6i = ''.join(name_6i)

   model_6i = ols(name_6i,data).fit()

   r_6i = model_6i.rsquared

   vif_6i = 1/(1-r_6i)

   sse_6i = model_6i.ssr

   cp_6i = sse_6i/sigma_square-size[0]+2*(6+1)

   print("%.60s  Cp = %.8s   VIF = %.8s" % (item,cp_6i,vif_6i))


for item in itertools.combinations(factors,5):
```

```
    P = '+'
    L = '~'
    name_5i = ('MDRI',L,item[0],P,item[1],P,item[2],P,item[3],P,item[4])
    name_5i = ''.join(name_5i)
    model_5i = ols(name_5i,data).fit()
    r_5i = model_5i.rsquared
    vif_5i = 1/(1-r_5i)
    sse_5i = model_5i.ssr
    cp_5i = sse_5i/sigma_square-size[0]+2*(5+1)
    print("%.60s  Cp = %.8s   VIF = %.8s" % (item,cp_5i,vif_5i))

for item in itertools.combinations(factors,4):
    P = '+'
    L = '~'
    name_4i = ('MDRI',L,item[0],P,item[1],P,item[2],P,item[3])
    name_4i = ''.join(name_4i)
    model_4i = ols(name_4i,data).fit()
    r_4i = model_4i.rsquared
    vif_4i = 1/(1-r_4i)
    sse_4i = model_4i.ssr
    cp_4i = sse_4i/sigma_square-size[0]+2*(4+1)
    print("%.60s  Cp = %.8s   VIF = %.8s" % (item,cp_4i,vif_4i))


for item in itertools.combinations(factors,3):
    P = '+'
    L = '~'
```

```
    name_3i = ('MDRI',L,item[0],P,item[1],P,item[2])
    name_3i = ''.join(name_3i)
    model_3i = ols(name_3i,data).fit()
    r_3i = model_3i.rsquared
    vif_3i = 1/(1-r_3i)
    sse_3i = model_3i.ssr
    cp_3i = sse_3i/sigma_square-size[0]+2*(3+1)
    print("%.60s  Cp = %.8s   VIF = %.8s" % (item,cp_3i,vif_3i))


for item in itertools.combinations(factors,2):
    P = '+'
    L = '~'
    name_2i = ('MDRI',L,item[0],P,item[1])
    name_2i = ''.join(name_2i)
    model_2i = ols(name_2i,data).fit()
    r_2i = model_2i.rsquared
    vif_2i = 1/(1-r_2i)
    sse_2i = model_2i.ssr
    cp_2i = sse_2i/sigma_square-size[0]+2*(2+1)
    print("%.60s  Cp = %.8s   VIF = %.8s" % (item,cp_2i,vif_2i))


for item in itertools.combinations(factors,1):
    P = '+'
    L = '~'
    name_1i = ('MDRI',L,item[0])
```

```
    name_1i = ''.join(name_1i)
    model_1i = ols(name_1i,data).fit()
    r_1i = model_1i.rsquared
    vif_1i = 1/(1-r_1i)
    sse_1i = model_1i.ssr
    cp_1i = sse_1i/sigma_square-size[0]+2*(1+1)
    print("%.60s  Cp = %.8s   VIF = %.8s" % (item,cp_1i,vif_1i))


test_class = OLSInfluence(model)
cook = test_class.cooks_distance[0]
np.savetxt('cooks.txt',cook)
```

Regression Tree:

```
clc
clear
rng(100,'twister')

data = readtable('dataset.csv');
% data = csvread('dataset.csv','mode');
y = data(:,2);
x = data(:,3:end);
n = 8;

Mdr1 = fitrtree(x,y,'Kfold',n);
mse1  = kfoldLoss(Mdr1);

x2 = data(:,4:end);
Mdr2 = fitrtree(x2,y,'Kfold',n);
mse2  = kfoldLoss(Mdr2);

Mdr3 = fitrtree(x,y);
view(Mdr3,'Mode','graph')
```